

JADAVPUR UNIVERSITY
DEPARTMENT OF INFORMATION TECHNOLOGY
M.E. INFORMATION TECHNOLOGY, FIRST YEAR, SECOND SEMESTER EXAM 2025
SUBJECT: ADVANCED DATABASES

Full Marks :100

Time : 3Hrs

	Answer any two questions 2 X 5 = 10 Q.1 i) Why is scalability a critical consideration in modern database systems, and how does it impact system performance and user experience as data volume and user load increase? CO1 : (10) ii) Compare and contrast vertical scaling and horizontal scaling in database systems. What are the strategic advantages and limitations of each approach? iii) How do distributed database architectures contribute to scalability, and what challenges must be addressed to maintain consistency, availability, and partition tolerance?															
	Answer any four questions 4 X 10 = 40 Q.2 i) a) Consider a scenario where a company wants to analyze sales data alongside customer demographic data to understand buying patterns. Show how Map Reduce can perform the relational join operation on these datasets. b) State the procedural differences between sharding and replication. 6+4 =10 CO2 : (40) ii) a) How the BASE properties of NoSQL databases do impact the design and scalability of distributed systems compared to the ACID properties of traditional relational databases? How can downtime be reduced in NoSQL databases? b) Explain the role and purpose of Thrift Server, Metastore and Managed tables in the context of Apache Hive? c) Where does the data of a Hive table get stored? 3+2+4+1 =10 iii) a) What are Column Families in HBase? Given a small dataset how would you organize the data into column families to optimize for frequent lookups of user details versus purchase history? <table><tr><th>UserID</th><th>Name</th><th>Email</th><th>Age</th><th>Purchase History</th></tr><tr><td>1</td><td>Alice</td><td>alice@email.com</td><td>30</td><td>Shoes, Shirt</td></tr><tr><td>2</td><td>Bob</td><td>bob@email.com</td><td>25</td><td>Laptop, Phone</td></tr></table> b) What is the Hierarchy of Tables in Apache HBase? c) Define compaction in HBase? (2+3)+3+2 =10	UserID	Name	Email	Age	Purchase History	1	Alice	alice@email.com	30	Shoes, Shirt	2	Bob	bob@email.com	25	Laptop, Phone
UserID	Name	Email	Age	Purchase History												
1	Alice	alice@email.com	30	Shoes, Shirt												
2	Bob	bob@email.com	25	Laptop, Phone												

[Turn over

	iv) a) How does Spark leverage lineage information to recover from faults during computation? b) Mention few Spark Control operations with their functionalities. c) How does the DAG scheduler optimize task scheduling for performance improvement? 3+3+4 =10 v) a) Give a scenario where RDDs' immutability and deterministic operations might become limiting constraints for real-time or iterative applications. b) A sliding window has a 15-minute duration and 5-minute slide interval. If an event occurs at 12:07 PM (Start time: 12.00), how many windows will it belong to? How does this affect aggregation performance? c) How does HBase's use of write-ahead logs (WAL) impact its write performance and durability compared to traditional relational databases? d) How does sorted maps work in HBase. e) How does lazy evaluation contribute to the optimization of Spark jobs? 2+2+2+2+2=10																																										
CO3 : (15)	Answer all questions (2X5)+3+2 =15 Q.3 a) Describe how the following data mining operations are applied and provide typical example for each. (i) Linear regression (ii) Boxplot analysis (iii) Correlation Analysis (iv) Clustering (v) Outlier Detection. b) What are the main steps involved in the Knowledge Discovery in Databases (KDD) process? c) What is data generalization in data mining, and how does it help in summarizing large datasets?																																										
CO4 :(20)	Answer any two questions 2 X 10 =20 Q.4 i) a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. I. Draw a schema diagram for the above data warehouse using any one of the schema of DW. II. Starting with the base cuboid [day; doctor; patient], what special OLAP operations should be performed in order to list the total fee collected by each doctor in 2004? b) What are the disadvantages of Fact Constellation schema? c) How many cuboids are possible from a 5-dimensional data? 6+2+2 =10 ii) a) Make use of the following table, a hypothetical employment statistics record for recent graduates to design a star schema containing employment statistics data for recent graduates based on your ideas. Think of other information that might be relevant in particular, like the warehouse contains data for multiple universities, multiple years and multiple majors as well. You must mention the steps of Data ware house design process. <table><tr><th>ID</th><th>major</th><th>avg. project score</th><th>avg. exam score</th><th>co-op?</th><th>employed?</th><th>Salary</th></tr><tr><td>1</td><td>Computer</td><td>87</td><td>75</td><td>Y</td><td>Y</td><td>60,000</td></tr><tr><td>2</td><td>History</td><td>?</td><td>92</td><td>N</td><td>N</td><td>?</td></tr><tr><td>3</td><td>Computer</td><td>77</td><td>95</td><td>N</td><td>Y</td><td>50,000</td></tr><tr><td>4</td><td>Engineering</td><td>97</td><td>65</td><td>N</td><td>N</td><td>0</td></tr><tr><td>5</td><td>Engineering</td><td>84</td><td>75</td><td>Y</td><td>Y</td><td>40,000</td></tr></table>	ID	major	avg. project score	avg. exam score	co-op?	employed?	Salary	1	Computer	87	75	Y	Y	60,000	2	History	?	92	N	N	?	3	Computer	77	95	N	Y	50,000	4	Engineering	97	65	N	N	0	5	Engineering	84	75	Y	Y	40,000
ID	major	avg. project score	avg. exam score	co-op?	employed?	Salary																																					
1	Computer	87	75	Y	Y	60,000																																					
2	History	?	92	N	N	?																																					
3	Computer	77	95	N	Y	50,000																																					
4	Engineering	97	65	N	N	0																																					
5	Engineering	84	75	Y	Y	40,000																																					

- b) Why the concept of data mart is important?
 c) What is a lattice of cuboids in a data warehouse, and how does it help to organize group-by aggregations in OLAP systems? 5+2+3 =10

iii)

- a) Consider the *Sales* relation given below. Write DMQL corresponding to the given queries.

- I. Find total sales.
 II. Find total sales item wise and color wise.

size		all			
		color			
item-name		dark	pastel	white	Total
	skirt	8	35	10	53
	dress	20	10	5	35
	shirt	14	7	28	49
	pant	20	2	5	27
Total		62	54	48	164

- b) What do you mean by Bitmap indexing in of OLAP Data? Illustrate with an example. 5+5 =10

Answer any one question

1X15 =15

Q.5

i)

- a) Explain with an example how uncontrolled privilege propagation in a DAC system can result in unauthorized access. What control mechanisms can be implemented to restrict this propagation?

CO5 : (15)

- b) Consider a case where a user X wants to illegally access the salaries of other employees but does not have the necessary privileges. Explain the type of attack X might attempt to perform, and describe how Mandatory Access Control policies can prevent the database from such attacks.

- c) What are the different types of threats to database security, and why is it important to identify and classify such threats? 5+6+4=15

ii)

- a) In a discretionary access control system, you want to authorize your secretary to fire employees (based on your instructions) and, before doing so, to check their employee records and current salaries. What specific privileges should you grant, and what would be the appropriate GRANT statements to assign those privileges?

- b) Why do statistical queries on small populations in a database pose a security risk to individual privacy? Clarify your answer with example.

- c) Mention some privileges available at account level. 5+6+4=15